Direct Preference Optimization

Tri Nguyen

Oregon State University

February 26, 2025

Alignment problem

- Human preference of a response $\boldsymbol{y}$ given a prompt $\boldsymbol{x}$ is measured by $r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$.
  - $r(\boldsymbol{x}, \boldsymbol{y}_1) > r(\boldsymbol{x}, \boldsymbol{y}_2)$ means $\boldsymbol{y}_1$ is more preferred than $\boldsymbol{y}_2$.
- Objective: given a trained language model $\pi_{\text{ref}}(\boldsymbol{y} \mid \boldsymbol{x})$, fine-tune it so that
  - The outputs are aligned with human preference, while
  - Retaining the original model's generation skill.

A realized objective function:

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \quad \underset{\boldsymbol{x} \sim \mathcal{D}, \boldsymbol{y} \sim \pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})}{\mathbb{E}} \left[r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x}, \boldsymbol{y}) \right] - \beta \underset{\boldsymbol{x} \sim \mathcal{D}}{\mathbb{E}} \left[D_{\text{kl}}(\pi_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) \parallel \pi^{\text{ref}}(\boldsymbol{y} \mid \boldsymbol{x})) \right] \tag{1}$$

Issues

1. $r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x}, \boldsymbol{y})$ is unknown.
2. Problem (1) is "hard" to optimize due to the involvement of $\boldsymbol{\theta}$ in $\boldsymbol{y} \sim \pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$ under expectation.

The RL from Human Feedback approach [Ziegler et al. 2019]

- Estimate the score function $r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x}, \boldsymbol{y})$
- Finetune the LLM model by optimizing the original objective function using the learned $r_{\boldsymbol{\phi}^\star}$.

Fixing Issue 1: Specifying Preference Model

In hope of learning $r_{\phi^\natural}(\boldsymbol{x}, \boldsymbol{y})$, we have to specify some model, and then obtain some samples.
Preference Bradley-Terry model:

- Given L items, item i has a score $s_i > 0$.
- It models a binary result of an event i *beats* j as a Bernoulli RV with parameter

$$\mathrm{Pr}(i \succ j) = \frac{s_i}{s_i + s_j}, \qquad \forall i, j \in [L].$$

In our LLM context,

$$\mathrm{Pr}(\boldsymbol{y}_1 \succ \boldsymbol{y}_2 \mid \boldsymbol{x}) = \frac{\exp(r_{\phi^\natural}(\boldsymbol{x}, \boldsymbol{y}_1))}{\exp(r_{\phi^\natural}(\boldsymbol{x}, \boldsymbol{y}_1)) + \exp(r_{\phi^\natural}(\boldsymbol{x}, \boldsymbol{y}_2))} = \sigma\big(r_{\phi^\natural}(\boldsymbol{x}, \boldsymbol{y}_2) - r_{\phi^\natural}(\boldsymbol{x}, \boldsymbol{y}_1)\big).$$

Under this model, the MLE objective is [Ziegler et al. 2019]

$$\underset{\boldsymbol{\phi}}{\text{minimize}} \quad \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2 \sim \mathcal{D}} \left[I[\boldsymbol{y}_1 \succ \boldsymbol{y}_2] \sigma\big(r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_2) - r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_1)\big) + I[\boldsymbol{y}_2 \succ \boldsymbol{y}_1] \sigma\big(r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_2) - r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_1)\big) \right],$$

But there is no guarantee of learning the true $r_{\phi^\natural}$.

Fixing Issue 2:

Now we have learned $r_{\boldsymbol{\phi}^\star}$, the objective is

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\left[r_{\boldsymbol{\phi}^\star}(\boldsymbol{x},\boldsymbol{y})\right]-\beta\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[D_{\mathsf{kl}}(\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x})\parallel\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x}))\right]\\
&=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\Big[r_{\boldsymbol{\phi}^\star}(\boldsymbol{x},\boldsymbol{y})-\beta(\log(\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x}))-\log(\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x})))\Big]\\
&=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})\right].
\end{aligned}
$$

This is a standard objective used in RL (policy gradient), hence can be solved using off-the-shelf tools such as PPO.

A new approach

Rafael Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *arXiv preprint arXiv:2305.18290* [2023]

$$\underset{\pi_{\boldsymbol{\theta}}}{\text{maximize}} \quad \underset{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}{\mathbb{E}}\left[r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})\right] - \underset{\boldsymbol{x}\sim\mathcal{D}}{\mathbb{E}}\left[D_{\mathsf{kl}}(\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x}) \,\|\, \pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x}))\right] \tag{2}$$

This problem has "closed-form" solution:

$$\pi^\star(\boldsymbol{y}\mid\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x})\exp\left(\frac{1}{\beta}r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})\right)$$

Note that RL people already known this, but this result is not very helpful due to the intractability of $Z(\boldsymbol{x})$.

Proof of optimal policy

$$
\begin{aligned}
\arg\max_{\pi_{\boldsymbol{\theta}}} \textsf{Objective} &= \arg\max_{\pi_{\boldsymbol{\theta}}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})} \left[r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})\right] - \beta \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}} \left[D_{\mathsf{kl}}(\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x}) \parallel \pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x}))\right] \\
&= \arg\max_{\pi_{\boldsymbol{\theta}}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})} \left[r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y}) - \beta \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x})}\right] \\
&= \arg\min_{\pi_{\boldsymbol{\theta}}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})} \left[\log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x})} - \frac{1}{\beta} r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})\right] \\
&= \arg\min_{\pi_{\boldsymbol{\theta}}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})} \left[\log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x})}{\frac{1}{Z(\boldsymbol{x})}\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x}) \exp\left(r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})/\beta\right)} - \log Z(\boldsymbol{x})\right] \\
&= \arg\min_{\pi_{\boldsymbol{\theta}}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})} \left[\log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}\mid\boldsymbol{x})}{\frac{1}{Z(\boldsymbol{x})}\pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x}) \exp\left(r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})/\beta\right)}\right],
\end{aligned}
$$

where $\frac{1}{Z(\boldsymbol{x})} = \sum_{\boldsymbol{y}} \pi_{\mathsf{ref}}(\boldsymbol{y}\mid\boldsymbol{x}) \exp \frac{r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x},\boldsymbol{y})}{\beta}$.

And therefore, the optimal value is 0 and optimal solution is

$$\pi^\star(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\pi_{\mathsf{ref}}(\boldsymbol{y} \mid \boldsymbol{x}) \exp \frac{r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x}, \boldsymbol{y})}{\beta}.$$

Now we can express the unknown score function $r()$ in terms of optimal solution $\pi^\star$, hence allow us to reduce the unknown to only $\pi^\star$.

$$r_{\boldsymbol{\phi}^\natural}(\boldsymbol{x}, \boldsymbol{y}) = \beta \log \frac{\pi^\star(\boldsymbol{y} \mid \boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y} \mid \boldsymbol{x})} + \beta \log Z(\boldsymbol{x})$$

Then with the preference model, we can derive the MLE objective to find that optimal $\pi^\star$.

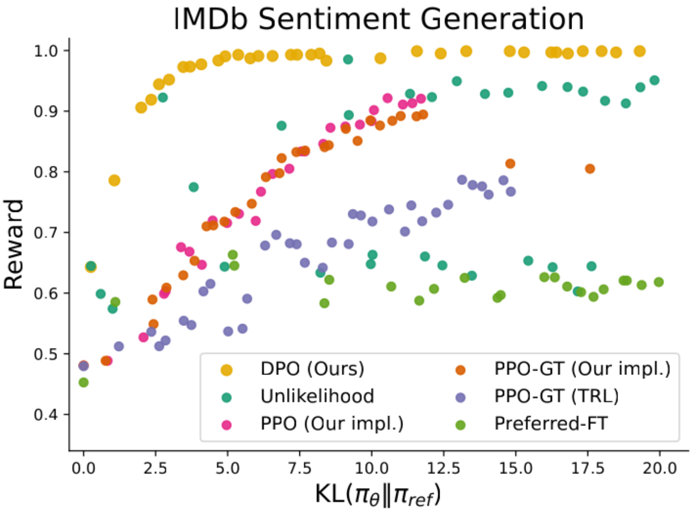
- Under the Bradley-Terry model, observing dataset $[(\boldsymbol{x}_i, \boldsymbol{y}_{1i}, \boldsymbol{y}_{2i})]_1^n$, the MLE objective is

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2 \sim \mathcal{D}} \left[I[\boldsymbol{y}_1 \succ \boldsymbol{y}_2] \sigma\big(r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_2) - r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_1)\big) + I[\boldsymbol{y}_2 \succ \boldsymbol{y}_1] \sigma\big(r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_2) - r_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{y}_1)\big) \right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2 \sim \mathcal{D}} \Bigg[I[\boldsymbol{y}_1 \succ \boldsymbol{y}_2] \sigma\Bigg(\beta \log \frac{\pi_{\boldsymbol{\phi}}(\boldsymbol{y}_1 \mid \boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_1 \mid \boldsymbol{x})} - \beta \log \frac{\pi_{\boldsymbol{\phi}}(\boldsymbol{y}_2 \mid \boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_2 \mid \boldsymbol{x})}\Bigg) \\
&\qquad + I[\boldsymbol{y}_2 \succ \boldsymbol{y}_1] \sigma\Bigg(\beta \log \frac{\pi_{\boldsymbol{\phi}}(\boldsymbol{y}_2 \mid \boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_2 \mid \boldsymbol{x})} - \beta \log \frac{\pi_{\boldsymbol{\phi}}(\boldsymbol{y}_1 \mid \boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_1 \mid \boldsymbol{x})}\Bigg). \Bigg]
\end{aligned}$$

In other words, we are parameterizing the unknown score function $r(\boldsymbol{x}, \boldsymbol{y}) = \log \pi_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) - \log \pi_{\text{ref}}(\boldsymbol{x}, \boldsymbol{y})$ to guarantee that the optimal solution of problem (1) is $\pi_{\boldsymbol{\theta}}$.

Control setting

We want to finetune a LM model such that it always produce positive reviews.



Control setting - My try

- Dataset: IMDB, $\sim$ 20k reviews
- True score function is given by a sentiment classifier (a pretrained large network)
- π_{ref}: Fine-tuning `gpt2-large` (1.4B params) on unlabeled IMDB
- For `PPO`, we provide the true score function.
- For `DPO`, given a prompt, we sample 4 responses for each prompt, and create 6 preference pairs.

Table: About an hour training for each method

	π_{ref}	π_{ppo}	π_{dpo}
Sentiment score	0.625	0.86	0.99
KL	0.	1.7	-26.6

Result on other tasks

Extensions

- Assuming preference pairs are noisy due to annotator's imperfection,

$$\begin{aligned} z &\sim \mathsf{Bern}(\sigma(r(\boldsymbol{x}, \boldsymbol{y}_1) - r(\boldsymbol{x}, \boldsymbol{y}_2))) \\ \ell &\sim \mathsf{Pr}(\ell' \mid z) \end{aligned}$$

- In [Christiano et al. 2017], some pairs annotations are just uniformed selected $\Rightarrow$ outliers.
- Instead of pairwise preferences, we can consider a best-choice preferences: Given a prompt $\boldsymbol{x}$ and L responses, the label is the best response. [Ziegler et al. 2019].
- Assuming existence of score function might not hold in general
- What about $D_{\mathsf{kl}}(\pi_{\mathsf{ref}} \parallel \pi_{\boldsymbol{\theta}})$

Preference Optimization with the Pairwise Cringe Loss

Jing Xu et al. "Some things are more cringe than others: Preference optimization with the pairwise cringe loss". In: *arXiv preprint arXiv:2312.16682* [2023] Alignment samples can be in different forms:

- Supervised setting: $(\boldsymbol{x}, \boldsymbol{y}\ )$
- Binary feedback: $(\boldsymbol{x}^+, \boldsymbol{y}^+, \boldsymbol{x}^-, \boldsymbol{y}^-\ )$
- Binary preference: $(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2\ )$

Cringe loss is originally applied to Binary feedback data:

$$\mathcal{L}_{\textsf{BIN}}(\boldsymbol{x}^-, \boldsymbol{y}^-, \boldsymbol{x}^+, \boldsymbol{y}^+) = \mathcal{L}_{\textsf{CE}} + \mathcal{L}_{\textsf{Cr}}$$
$$\mathcal{L}_{\textsf{CE}}(\boldsymbol{x}^+, \boldsymbol{y}^+) = -\log \textsf{Pr}(\boldsymbol{y}^+ \mid \boldsymbol{x}^+)$$
$$\mathcal{L}_{\textsf{Cr}}(\boldsymbol{x}^-, \boldsymbol{y}^-) = -\log \sum_t \log \frac{\exp(s_t^*)}{\exp(s_t^*) + \exp(s_t[y_t^-])},$$

where we feed the prompt $\boldsymbol{x}^-$ to the model, and ask it to generate an output of length T:

- At the t-th token, we select top k tokens according model's prob output $s_t^1, \ldots, s_t^k$.
- Normalizing probability over these tokens by applying softmax function.
- Sample an index $z \sim \textsf{Categorical}(s_t^1, \ldots, s_t^k), z \in [k]$.
- $s_t^* = s_t^z$.

Apply Cringe Loss to Pairwise Preference data

They propose to use the following loss on pairwise preference data

$$\mathcal{L}_{\mathsf{Pair}}(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2) = g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)\mathcal{L}_{\mathsf{BIN}}(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{x}, \boldsymbol{y}_2),$$

where

$$\begin{aligned} g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2) &= \sigma(b - M(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)), \\ M(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2) &= \log \mathsf{Pr}(\boldsymbol{y}_1 \mid \boldsymbol{x}) - \log \mathsf{Pr}(\boldsymbol{y}_2 \mid \boldsymbol{x}). \end{aligned}$$

Result

Table 1: AlpacaFarm evaluation results (LLM evaluation), using human preference data and reward model (where applicable) for training. (*=average of 3 seeds). [1]PPO with human preferences was trained by Dubois et al. (2023); we just evaluated the model.

Method	Win rate (%)
Results reported by Dubois et al. (2023)	
Llama 7B	11.3
SFT 10k	36.7
SFT 52k	39.2
Experiments reported in this paper:	
Binary Cringe	47.7*
PPO[1]	48.5*
DPO	50.2*
Pairwise Cringe	54.7*

Figure: Image

[1] Paul F Christiano et al. "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30 (2017).

[2] Rafael Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *arXiv preprint arXiv:2305.18290* (2023).

[3] Jing Xu et al. "Some things are more cringe than others: Preference optimization with the pairwise cringe loss". In: *arXiv preprint arXiv:2312.16682* (2023).

[4] Daniel M Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).